

Identifying transcription factor binding sites by cross-species comparison

Lee Ann McCue
Bioinformatics Group
Wadsworth Center
mccue@wadsworth.org

E. coli transcription regulatory network

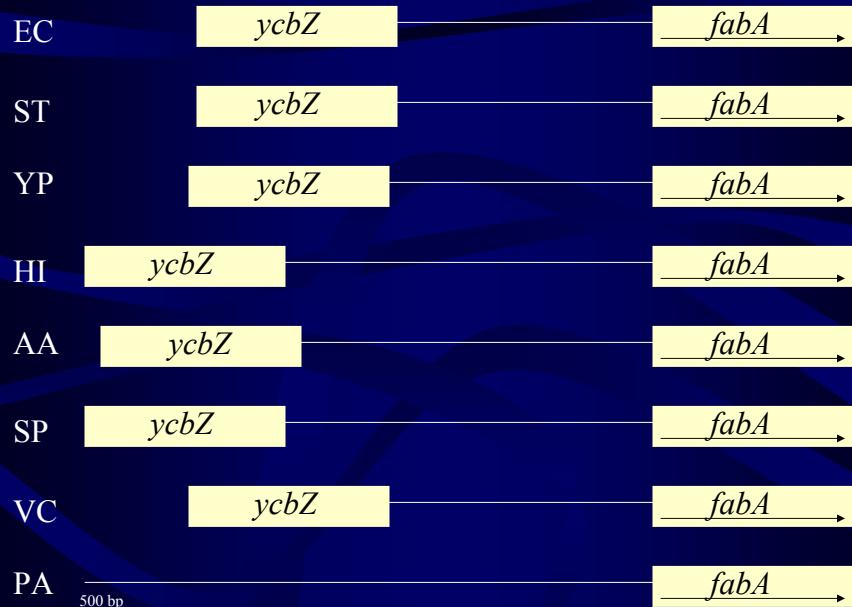
- Identify the transcription factors (TF)
- Identify the TF binding sites
- Connect the sites with the cognate TFs
- Associate activity with each connection
 - activation or repression
 - affinity of binding
- Identify the signals controlling the TFs

Phylogenetic footprinting

- Comparative genomics approach

(assumes similar regulation of orthologous genes cross-species)

- identify the orthologous genes in related species
- use an alignment method to identify the motif
- search the genome for more instances of the motif



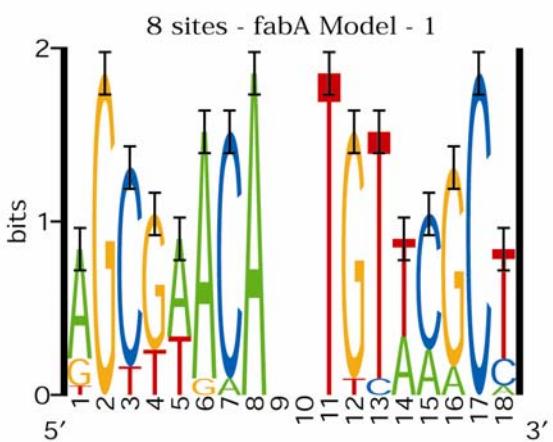
Gibbs Sampling Strategy

Lawrence et al. (1993) *Science* 262:208-214
Neuwald, Liu, & Lawrence (1995) *Protein Sci* 4:1618-1632

- specify palindromic models
Lawrence & Reilly (1990) *Proteins* 7:41-51
- site width = 16 or 17 bases, may fragment to 24
Liu, Neuwald & Lawrence (1995) *JASA* 90:1156-1170
- incorporate a distribution of spacing model
- incorporate position-specific background model
Liu & Lawrence (1999) *Bioinformatics* 15:38-52
- configured to detect up to 2 sites per sequence
- multiple runs, order results by maximum *a posteriori* probability (MAP) value

Validation of predictions

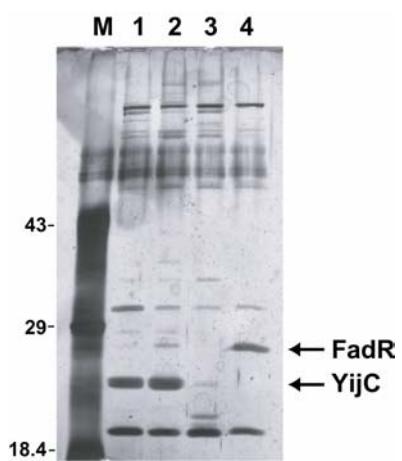
- Study Set:
genes with experimentally verified TF binding sites in the promoter
- most probable motif predictions:
 - 81% are known sites, given ≥ 3 species in the data
- among the remaining predictions:
 - false positives?
 - regulation not conserved cross-species?
 - some are real, but previously undocumented sites - *fabA*



fabA 5' AGCTAACACGTGTACGCT 3'

fabB 5' AGCGTACACTTGTACGCC 3'

yqfA 5' AGTGAACACCTGTTAGCT 3'



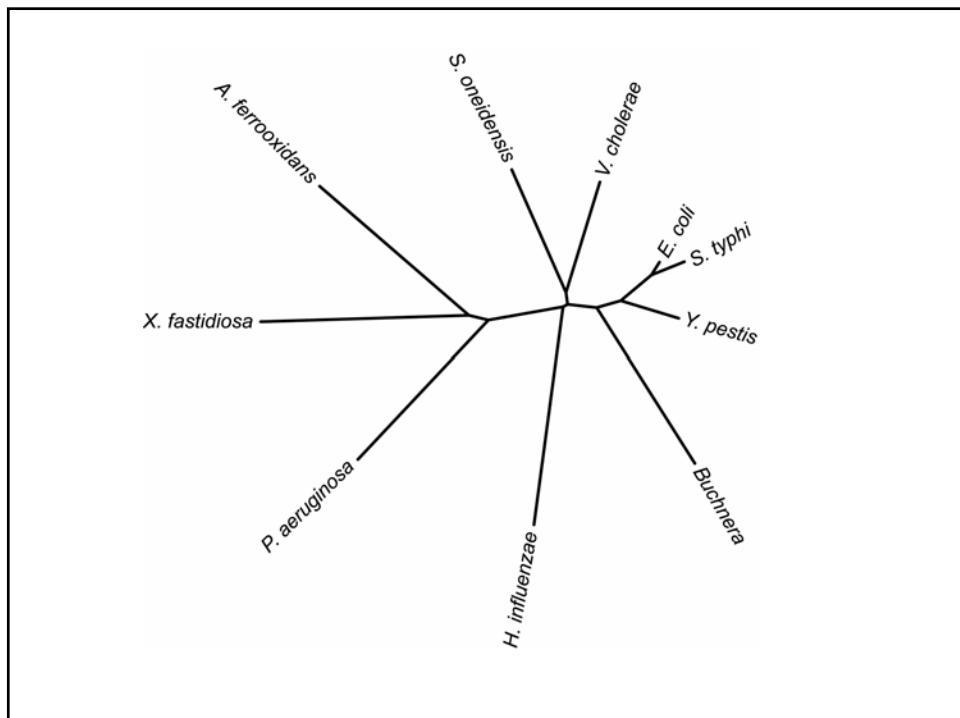
E. coli transcription regulatory network

- Identify the transcription factors (TF)
- Identify the TF binding sites
- Connect the sites with the cognate TFs
- Associate activity with each connection
 - activation or repression
 - affinity of binding
- Identify the signals controlling the TFs

FAQs

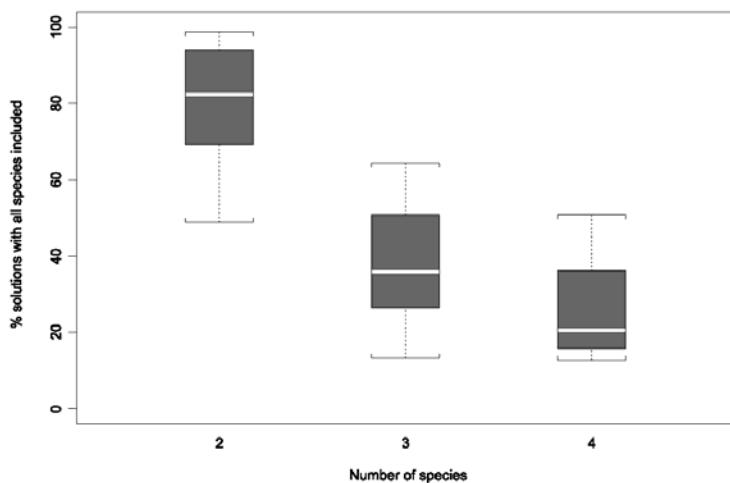
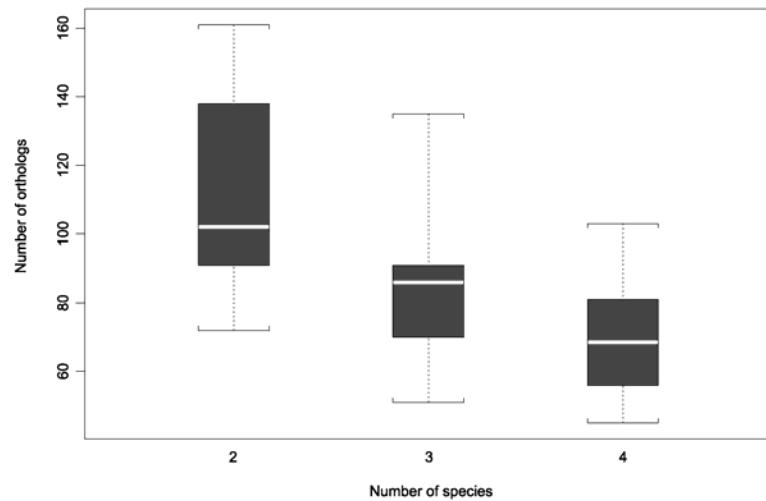
- How many species do you need?
- How do you choose species?
- How do you prioritize predictions for affinity purification?
(What's statistically significant?)

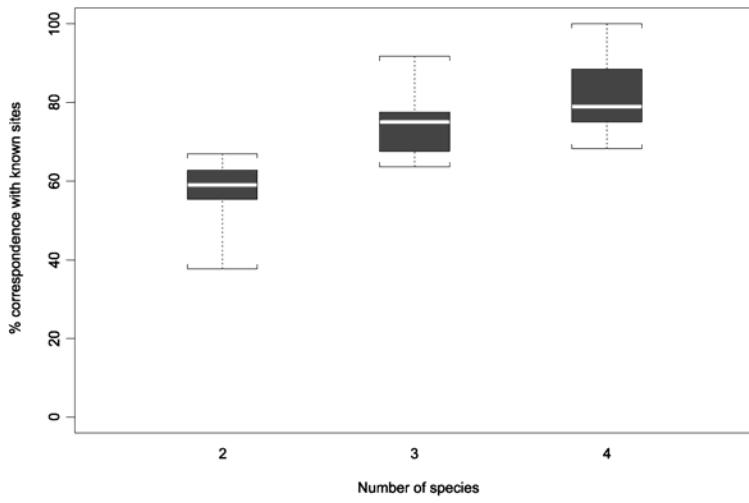
<u>Species</u>	<u>Habitat(s)</u>	<u>Genome size</u>
<i>Escherichia coli</i>	intestine	4.64 Mbp
<i>Salmonella typhi</i>	intestine & bloodstream	4.81 Mbp
<i>Yersinia pestis</i>	bloodstream	4.65 Mbp
<i>Buchnera</i> sp. APS	aphid bacteriocytes	0.64 Mbp
<i>Haemophilus influenzae</i>	nasopharynx	1.83 Mbp
<i>Vibrio cholerae</i>	small bowel	4.03 Mbp
<i>Shewanella oneidensis</i>	marine sediments	4.50 Mbp
<i>Pseudomonas aeruginosa</i>	multiple human sites	6.26 Mbp
<i>Acidithiobacillus ferrooxidans</i>	acidic water/soil	2.90 Mbp
<i>Xylella fastidiosa</i>	xylem of plant	2.68 Mbp



<u>Species</u>	<u>Study Set orthologs</u>	<u>TF orthologs</u>
<i>E. coli</i>	166	48
<i>S. typhi</i>	161	46
<i>Y. pestis</i>	138	40
<i>Buchnera</i>	27	1
<i>H. influenzae</i>	72	21
<i>V. cholerae</i>	112	31
<i>S. oneidensis</i>	91	24
<i>P. aeruginosa</i>	92	25
<i>A. ferrooxidans</i>	46	11
<i>X. fastidiosa</i>	50	7

<u>Species</u>	<u>Data sets</u>	<u>Predictions w/all species</u>	<u>Known sites</u>
EC-ST	161	98.8% (159)	54.7% (87)
EC-YP	138	94.2% (130)	66.2% (86)
EC-VC	112	83.0% (93)	62.4% (58)
EC-HI	72	81.9% (59)	62.7% (37)
EC-SO	91	69.2% (63)	55.6% (35)
EC-PA	92	48.9% (45)	35.6% (16)
EC-BU	27	85.2% (23)	39.1% (9)
EC-AF	46	58.7% (27)	48.1% (13)
EC-XF	50	54.0% (27)	33.3% (9)
EC-VC-HI	65	50.8% (33)	75.8% (25)
EC-VC-SO	87	46.0% (40)	67.5% (27)
EC-YP-PA	86	19.8% (17)	70.6% (12)
EC-ST-YP-VC	103	42.7% (44)	86.4% (38)



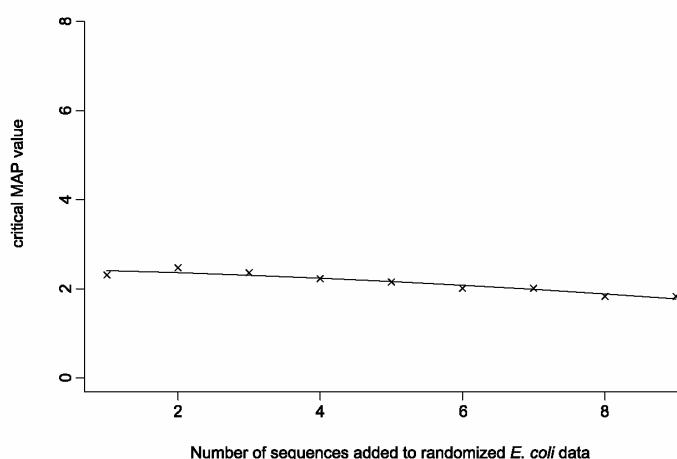


Using all available data (7 species)

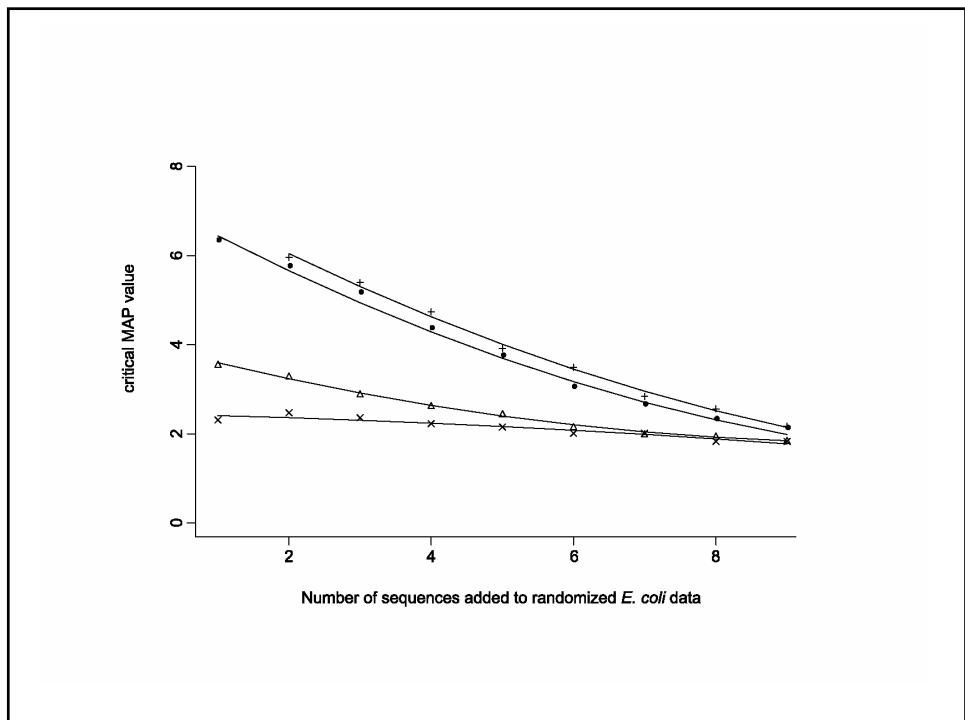
<u>166 gene set</u>	<u>number</u>	<u>known sites</u>
predictions \geq 2 species	163	69.3% (113)
predictions \geq 3 species	117	75.2% (88)
predictions \geq 4 species	85	75.3% (64)

Statistical significance of the predictions

- simulations with randomized data
- predictions sorted by “average MAP” value
MAP / number of sites in prediction
- determine the 95% quantile
→ critical MAP value



Species	Study Set orthologs	Sequence data, % identity
<i>E. coli</i>	166	100
<i>S. typhi</i>	161	70.0 ± 12.4
<i>Y. pestis</i>	138	48.1 ± 7.9
<i>H. influenzae</i>	72	41.5 ± 4.9
<i>V. cholerae</i>	112	41.6 ± 6.0
<i>S. oneidensis</i>	91	41.1 ± 5.0
<i>P. aeruginosa</i>	92	37.7 ± 3.7



Using all available data (7 species)

- 166 gene set, predictions with ≥ 3 species
 - 117 genes
 - 66 have an avgMAP \geq critical value (56.4%)
 - 53 have an avgMAP \geq critical value AND are a known *E. coli* site
 - 35 have an avgMAP $<$ critical value AND are a known *E. coli* site

Genome scale phylogenetic footprinting

- 2086 *E. coli* genes
(at least one ortholog and ≥ 50 bp upstream)
- 587 of these have avgMAP \geq critical value
(but this counts only the top solution for each gene!)
- 802 total predictions with avgMAP \geq critical value

Conclusions

- Species selection
 - simple species characteristics as predictors
 - number of orthologs
- Number of species → 3 or 4!
- Statistical significance
 - critical MAP value as a measure
 - correlated sequence data impacts critical value

Genome Sequence Data

- GenBank
 - *Escherichia coli* (Blattner *et al.* 1997)
 - *Buchnera* sp. APS (Shigenogu *et al.* 2000)
 - *Pseudomonas aeruginosa* (Stover *et al.* 2000)
 - *Xylella fastidiosa* (Simpson *et al.* 2000)
- The Institute for Genomic Research
 - *Haemophilus influenzae* (Fleischmann *et al.* 1995)
 - *Shewanella oneidensis*
 - *Acidithiobacillus ferrooxidans*
 - *Vibrio cholerae* (Heidelberg *et al.* 2000)
- Sanger Centre
 - *Salmonella typhi*
 - *Yersinia pestis* (Parkhill *et al.* 2001)

The Wadsworth Center

Bioinformatics

Chip Lawrence
Bill Thompson
Steve Carmack
Ivan Auger
Mike Palumbo
Linda Mayerhofer

Molecular Biology

Vicky Derbyshire
Mike Ryan
Bill Albano

Mass Spectrometry

Charles Hauer
Bob Stack

Harvard University

Jun Liu